IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 43-47
www.iosrjournals.org

Sentiment Analysis of Movie Ratings System

Sagar Chavan[1], Akash Morwal[2,]Shivam Patanwala[3,]Prachi Janrao[4]
[1](Department of Computer Engineering, Thakur College of Engineering and Technology, India)
[2](Department of Computer Engineering, Thakur College of Engineering and Technology, India)
[3](Department of Computer Engineering, Thakur College of Engineering and Technology, India)
[4](Department of Computer Engineering, Thakur College of Engineering and Technology, India)

Abstract: *Sentiment analysis also term as refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. In recent years, Opinion mining is a hotspot in the field of natural language processing, and it is also a challenging problem Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. Movie reviews are an important way to gauge the performance of a movie. The objective of this paper is to extract features from the product reviews and classify reviews into positive, negative. . In this project we aim to use Sentiment Analysis on a set of movie reviews which is given by reviewers and then try to understand what the overall reaction to the movie was according to them, i.e. if they liked the movie or they hated it. We aim to use the relationships of the words in the review to predict the overall polarity of the review.*

Keywords: *Movie Review Mining, Pre-processing, Sentiment Analysis, Stemming*

I. Introduction

The growth of content in the Internet in recent years has made a huge quantity of information available. This information is presented in different formats such as posts, news articles, comments, and reviews. Reviews, comments and opinions of the people play an important role in determining whether a given population is satisfied with a product or a service or in judging their response to specific event. Data consisting of such reviews or opinions has a very high potential for knowledge discovery. One of the basic tasks in SA is to predict the polarity of a given sentence, to find out if it expresses a positive or negative feeling about a certain topic [14] Sentiment analysis isan interdisciplinary field that crosses natural language processing, artificial intelligence, and text mining. Since most opinions are available to us in the text format and its processing is easier than other formats, sentimentanalysis has emerged as a subfield of text mining [3].Sentiment analysis is used in different domains, ranging from analyzing tweets to comments on a particular website. With good amount of research Sentiment Analysis, they can even be used on social issues. This paper focuses on sentiment classification in Movie Reviews domains.

Text mining process is similar to data mining, except, the data mining tools are made to handle structured data whereas text mining can be used to handle unstructured or semi-structured data sets such as emails HTML files and full text documents etc. [1].The unstructured data is something which usually refers to information which doesn't reside in a traditional row-column database. Semi-Structured data is the data which is neither raw data, nor typed data as to that of a conventional database system

II. Related Work

Most of the works on sentiment analysis have been done at the document level. Paper by A.Jeyapriya et al on "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm" [2] discusses phrase- level opinion mining which performs finer grained analysis and directly look at the opinion in the online reviews which is used to extract important aspects of an item and to predict the orientation. It makes use of Machine Learning to find the sentiment orientation.

Kang Wu et al., [4] focus on sentiment analysis of topical Chinese microblogs. In this paper one of the popular microblog of China is considered i.e. Sina WeiBo. User of WeiBo who writes their messages which usually have various sentences, messages length is up to 140 Chinese Microblog have several sentences, which allow bloggers to share their opinion. The conducted Study shows that Chinese people show their sentiments in indirect way. The proposed model first, analyzed the Chinese Microblogs which tells the opinion of user, and analyzes the features of single sentence. Second, in order to optimize the result of sentiment classification we use sentence relationship. V.K. Singh et al., [6] presents a new method which is feature based domain specific heuristic for aspect level sentiment classification for reviews of movie. It analyzes the movie reviews and then on each aspect it analyses and assigns a sentiment label. After that it aggregate the scores on each aspect from

various different reviews and then on all parameters, a net sentiment profileis generated. In this the author is also using SentiWordNet with linguistic features such as adverbs, adjectives and verbs. As Sentiment Analysis is mostly covered for marketing research purposes in order to find out the generalized opinion of the products or services offered the Paper "Sentiment Analysis of Social Issues" by Mostafa Karamibekret al [9] suggested that sentiment analysis can also be applied to social issues. Based on our findings, we propose an approach to take into account the role of verb as the most important term in expressing opinions regarding the social issues. Experimental results show that considering verbs is very necessary and it also improves the performance of sentiment analysis.

Yu Huangfu et al in paper "An Improved Sentiment Analysis Algorithm for Chinese News" [8] proposed their own new method known as Improved Sentiment Analysis (ISA) which takes into consideration text and tittle sentiment analysis. Experimental data shows that the proposed method improves the accuracy of news sentiment analysis.V.K. Singh et al., [10] used a SentiWordNet based approach with two linguistic feature. The SentiWordNet approach considered in this case for document level sentiment classification of movie reviews and blog posts is implemented in this paper and performance evaluation is also made on the same. With different variations of linguistic features, threshold for aggregation and scoring schemes, SentiWordNet is implemented. Here it is using two approaches: i) SVM (Support Vector Machine) and Naïve Bayes forthe classification of sentiments. Also papers Tirath Prasad et al [11] and Purtata Bhoir et al [12] performed sentiment analysis with the help of SentiWordNetin order to find the overall polarity. Also Pre-processing is missing to quite an extent in these papers which is essential in improving the accuracy.

III. Proposed Methodology

This section describes the different preprocessing modules that have been used in our project. All of them are built in Python. The pipeline of the project is organized in the following way. In order to perform sentiment analysis, data have to be prepared in order to obtain a data set – namely, the training set. The training sets are subject to preprocessing techniques mentioned in the work. Then, such a data set is involved in the learning step, which uses Machine Learning [ML] algorithm and yields a trained classifier. After training the classifier it has to be tested on a different data set – namely the test set. Figure 2 shows the various steps for training a classifier to perform sentiment analysis

3.1 Preprocessing Phases:

3.1.1Basic Operation and Cleaning: This first module manages basic cleaning operations, which consist in removing unimportant or disturbing elements for the next phases of analysis and in the normalization of some misspelled words. In order to provide only significant information, in general a clean review should not contain URLs and hashtags (i.e. #happy). Furthermore, tabs and line breaks should be replaced with a blank and quotation marks with apexes. After this step, all the punctuation is removed, except for apexes, because they are part of grammar constructs such as the genitive.The next operation is to remove the vowels repeated in sequence at least three times, because by doing so the words are normalized: for example, two words written in a different way (i.e. cooooool and cool) will become equals.

The last step is to convert many types of emoticons into tags that express their sentiment (i.e. :) → smile happy). The list of emoticons is taken from Wikipedia. Finally, all the text is converted to lower case, and extra blank spaces are removed. All the operations in this module are executed to try to make the text uniform. This is important because during the classification process, features are chosen only when they exceed a certain frequency in the data set. Therefore, after the basic preprocessing operations, having different words written in the same way helps the classification.

3.1.2Dictionary: Here we make use of the external python library PyEnchant, which provides a set of functions for the detection and correction of misspelled words using a dictionary. It also allows us to substitute slang with its formal meaning (i.e., l8 → late), using a list. It also allows us to replace insults with the tag "bad word". The reason for the use of these functions is the same as for the basic preprocessing operation, i.e. to reduce the noise in text and improve the overall classification performances.

3.1.3Negation: Dealing with negations (like "not good") is a critical step in Sentiment Analysis. A negation word can influence the tone of all the words around it, and ignoring negations is one of the main causes of misclassification. In this phase, all negative constructs (can't, don't, isn't, never etc.) are replaced with "not". This technique helps classifier model to be enriched with a lot of negation constructs that would otherwise be excluded due to their low frequency as detailed in Fig. 2, Machine Learning algorithms need to work on data that is properly processed. This phase is a fundamental step in order for the whole system to obtain good results. Normally it includes methods for data cleaning and feature extraction and selection. A good overview of the steps and the most known algorithms for each step is explained in [13].

3.1.4 Stop Words: Stop words are a division of natural language. The reason that stop-words should be removed from a text is that they make the text look heavier. For it we have various stop word removing techniques [5] were referred. Removing stop words helps in reducing the dimensionality of term space. The most common words in text documents are articles, prepositions, etc that do not give the meaning to the documents. These are the wordswhich are treated as stop words. Example: the, in, a, an, with, etc. It is important to avoid having these words within the classifier model, because they can lead to a less accurate classification.

3.1.5 Stemming: Stemmers remove morphological affixes from words, leaving only the word stem.For example, the words connect, connected, connecting, connections all can be stemmed to the word "connect" [15].The purpose of this method is to remove various suffixes, to have accurately matching stems, to save time and memory space. This is illustrated in Fig. 3 Translation of morphological forms of a word to its stem is done assuming that the words are semantically related. There are two points are considered while using a stemmer:

- Words that do not have the same meaning should be kept separate
- Morphological forms of a word are assumed to have the same base meaning and hence it should be mapped to the same stem

These two rules are good and sufficient in language processing applications. For our Project we will be using Snowball Stemmer with the help of NLTK toolkit. It is an evolution of the original Porter Stemmer algorithm and is computationally fast and more efficient.

3.1.6 POS Tagging: The Part-Of-Speech of a word is a linguistic category which is defined by what is known as its syntactic or morphological behavior. Common POS categories in English grammar are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the task of labeling (or tagging) each word in a sentence with its appropriate part of speech. POS tagging is animportant phase of opinion mining, it is essential to determine the features and opinion words from the reviews. POS tagging can be done either manually or with the help of POS tagger tool. POS tagging of the reviews by human is time consuming. POS tagger is used to tag all the words of reviews. Stanford tagger is used to tag each word in an online review sentences. Every one sentence in customer reviews are tagged and stored in text file

IV. Algorithm

The classifier is made by using Naive-Bayes Multinomial (NBM) method, i.e., a ML algorithm that gives rise to a probabilistic classifier, which works on the basis of the Bayes Theorem, with the strong assumption that features are mutually independent. Let X = (x_1, . . . , x_n) be the feature vector of an instance in the data set, that is, a binary vector that takes into account the presence of a feature in that instance, and let C_1, . . . , C_k be the possible outputs (classes). The problem is to gain the posterior probability of having the class C_k as output, given the feature vector X, and given the prior probability p (C_k) for each class.

Thanks to the Bayes Theorem and the independence between features, the probability that needs to be estimated is the conditional p(X|C_k), and then a classifier is trained with a decision rule, such as the Maximum a Posteriori (MAP) rule. In summary, the probabilistic model of NBM can be expressed in terms of the following formula:

$$\boldsymbol{P(X/C_k)} = \frac{(\sum_i \boldsymbol{x_i})!}{\prod_i \boldsymbol{x_i}!} \prod_i \boldsymbol{p_{ki}^{x_i}} \qquad \text{…Eq (1)}$$

Where X = (x_i. . . x_n) is the feature vector, pi is the probability that the feature i appears, C_k is a class and p_{ki} is the probability that feature i occurs in the class C_k. Then, Information Gain (IG)is the algorithm used for feature selection [REF]. It evaluates the presence or absence of a feature in a document by measuring its probability of belonging to a class. The amount of information needed to exactly classify an instance D is defined recursively as follows:

$$\boldsymbol{Info_A(D)} = -\sum_{j=1}^{v} \frac{|\boldsymbol{D_j}|}{|\boldsymbol{D_i}|} . \boldsymbol{Info(D_j)} \qquad \text{… Eq (2)}$$

When the instance D is divided by some feature attribute A = {a_1. . . a_v} into sub-instances D_1, . . . , D_v.

V. Figures And Tables

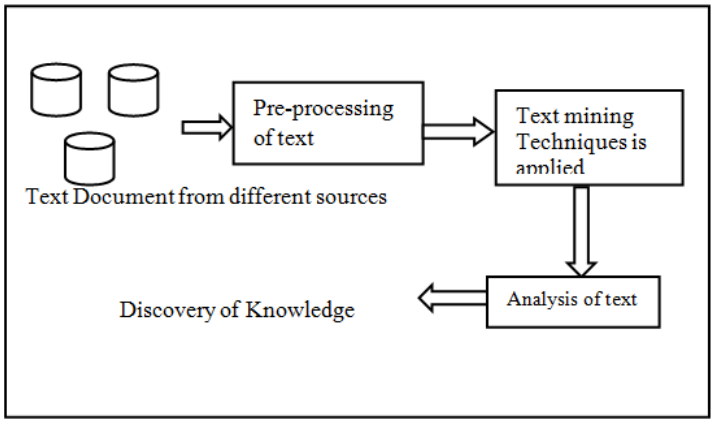

Figure 1. Text Mining Process

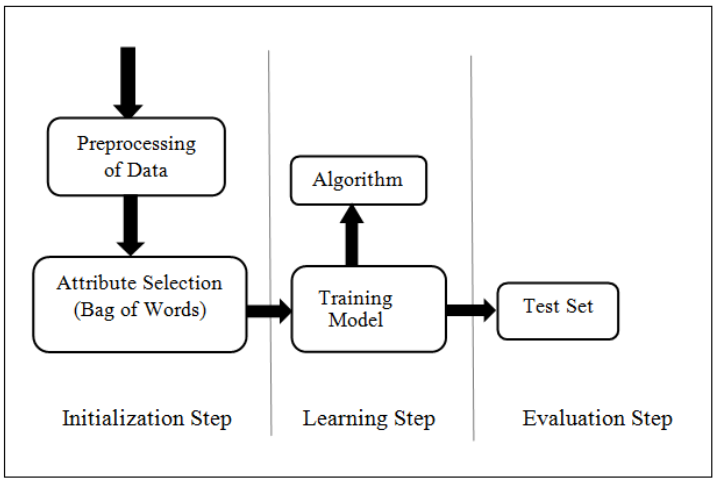

Fig. 2. Steps for training a classifier for sentiment analysis

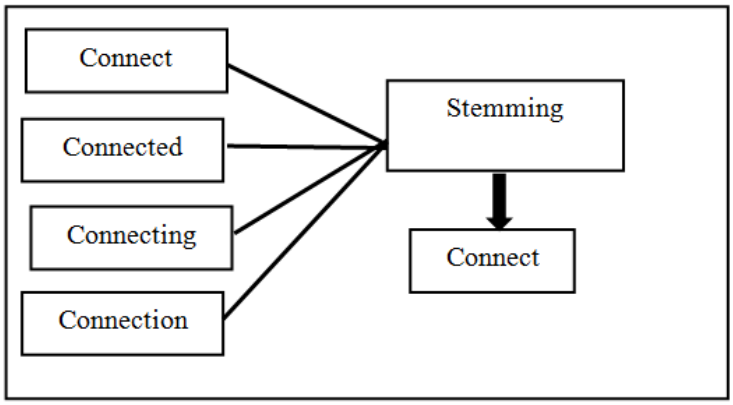

Fig 3 – Stemming Process

VI. Conclusion

Text preprocessing is an important phase in all relevant applications of data mining. In Sentiment Analysis, in particular, it is cited in almost all available research works. The Implementation of the project was carried out on data of movie reviews. Sentiment Analysis and Opinion mining has become a fascinating research area due to the availability of a huge volume of user-generated content in review sites, forums and blogs. Sentiment Analysis has applications in a variety of fields ranging from market research to decision making to advertising. With the help of Sentiment Analysis, companies can estimate the extent of product acceptance and can devise strategies to improve their product.

VII. Results And Discussion

```
Original Naive Bayes Algo accuracy percent: 74.84939759036145
Most Informative Features
              engrossing = True              pos : neg    =     20.0 : 1.0
                  boring = True              neg : pos    =     19.8 : 1.0
                 generic = True              neg : pos    =     16.6 : 1.0
                 routine = True              neg : pos    =     16.0 : 1.0
                mediocre = True              neg : pos    =     16.0 : 1.0
               inventive = True              pos : neg    =     14.7 : 1.0
                    flat = True              neg : pos    =     14.5 : 1.0
              refreshing = True              pos : neg    =     14.1 : 1.0
                    warm = True              pos : neg    =     12.0 : 1.0
               wonderful = True              pos : neg    =     11.6 : 1.0
                mindless = True              neg : pos    =     11.2 : 1.0
                    dull = True              neg : pos    =     11.0 : 1.0
                touching = True              pos : neg    =     10.8 : 1.0
           extraordinary = True              pos : neg    =     10.8 : 1.0
                  stupid = True              neg : pos    =     10.4 : 1.0
MNB_classifier accuracy percent: 73.04216867469879
BernoulliNB_classifier accuracy percent: 75.6024096385542
LogisticRegression_classifier accuracy percent: 73.19277108433735
LinearSVC_classifier accuracy percent: 71.83734939759037
SGDClassifier accuracy percent: 72.13855421686746
```

Fig 4 - Training Algorithm

```
This movie was awesome! The acting was great, plot was wonderful, and there were pythons...so yea!
('pos', 1.0)
This movie was utter junk. There were absolutely 0 pythons. I don't see what the point was at all. Horrible movie, 0/10
('neg', 1.0)
```

Fig 5 - Result of Sentiment Analysis

References

[1]. Vishal Gupta and Gurpreet S. Lehal, a Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[2]. A.Jeyapriya and C.S.Kanimozhi Selvi Extracting Aspects and Mining Opinions in Product Reviews Using Supervised LearningAlgorithm

[3]. V. P. H. Binali and W. Chen. A State Of The Art Opinion Mining and Its Application Domains. In IEEE International Conference On Industrial Technology, Pages 1–6, February 2009.

[4]. Su, Et Al, "Hidden Sentiment Association in Chinese Web Opinion Mining"

[5]. M.F. Porter, An Algorithm For Suffix Stripping, Program, Vol. 14, No. 3, Pp. 130-137, 1980.

[6]. V.K. Singh, R.Priyani, A. Uddin and P.Waila, "Sentiment Analysis of Movie Reviews a New Feature-Based Heuristic For Aspect-Level Sentiment Classification" 978-1-4673-5090- 7/13 IEEE, 2013

[7]. Ms. Anjali Ganesh Jivani, A Comparative Study Of Stemming Algorithms, Anjali Ganesh Jivani Et Al, Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938, ISSN: 2229-6093.

[8]. Yu Huangfu, Guoshi Wu, Yu Su, Jing Li, Pengfei Sun, And Jie Hu "An Improved Sentiment Analysis Algorithm For Chinese News"

[9]. Sentiment Analysis of Social Issues by Mostafa Karamibekr and Ali A. Khorbani

[10]. V.K. Singh, R.Priyani, A. Uddin and P.Waila,"Sentiment Analysis of Movie Reviews and Blog Posts Evaluating SentiWordNetwith Different Linguistic Features and Scoring Schemes", 3rd IEEE International Advance Computing Conference (Iacc), 978-1-4673- 4529-3/12, 2012.

[11]. Tirath Prasad Sahu, Sanjeev Ahuja Sentiment Analysis of Movie Reviews: A Study on Feature Selection & Classification Algorithms

[12]. Purtata Bhoir, Shilpa Kolte Sentiment Analysis of Movie Reviews Using Lexicon Approach

[13]. Data Pre-processing For Supervised Leaning S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas

[14]. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)

[15]. C.Ramasubramanian and R.Ramya, Effective Pre-Processing Activities in Text Mining Using Improved Porter"S Stemming Algorithm, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013, ISSN (Online): 2278-1021.